Procedia Computer Science

Volume 80, 2016, Pages 1822–1833

ICCS 2016. The International Conference on Computational Science

Bounded Support and Confidence over Evidential Databases

Ahmed Samet and Tien Tuan Dao

Sorbonne University, Université de technologie de Compiègne
CNRS, UMR 7338 Biomechanics and Bioengineering.
ahmed.samet@utc.fr tien-tuan.dao@utc.fr

Abstract
In this work, we propose a new definition of support and confidence measures based on interval representation. Moreover, a new algorithm, named EBS-Apriori, based on these bounded measures and several pruning strategies is developed. A new associative classifier, named WEvAC, based on fusion and weighting technique is implemented and tested. Experiments are conducted using several database benchmarks. Performance analysis showed a better prediction outcome for our proposed approach in comparison with several literature-based methods.

Keywords: Evidential database, Bounded support, Bounded confidence, Associative classifier.s

1 Introduction

Modern data acquisition is commonly characterized by the presence of uncertainty and imprecision leading to a new research challenge. When data mining techniques are applied to these data, their uncertainty has to be considered to obtain high quality results as well as to interpret prediction outcomes with more confidence. Therefore, several frameworks are used to represent the uncertainty and imprecision such as probabilities [1], fuzzy set theory [3] and more recently evidence theory [6]. The latter has led to the emergence of a new kind of database, named evidential database, that generalizes probabilistic and binary databases [12]. Thus, several basic concepts related to data mining domain has to be revised and this remains a challenging issue for the community. In particular, when dealing with evidential databases, several researches tackled the computing of the support and the confidence measures [1, 6, 9, 13]. The *support* represents the frequency of appearance of a pattern within a database. For fuzzy and even for evidential databases, several measures exist depending on the used strategy. For example in fuzzy data mining, several measures were introduced depending on the studied context and application such as [3, 7]. Thus, a main concern arises: which support measure should be chosen?. The same question could be asked for the *confidence* measure that computes how pertinent a rule is. Several measures have been proposed to compute the support and the confidence in evidential data mining. In [6], the authors proposed a belief-based measure of support. It relies

 Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2016

doi:10.1016/j.procs.2016.05.469

Transaction	Attribute A	Attribute B
	$m(A_1) = 0.7$	$m(B_1) = 0.4$
T1	$m(\Theta_A) = 0.3$	$m(B_2) = 0.2$
		$m(\Theta_B) = 0.4$
T2	$m(A_2) = 0.3$	$m(B_1) = 1$
	$m(\Theta_A) = 0.7$	

Table 1: Example of an evidential database $\mathcal{EDB}$

on a pessimistic estimation of the support. Another measure of support has been proposed relying on a probabilistic formulation [13]. Moreover, in [11], Samet et al. introduced a new measure of confidence for association rules based on support measure. In fact, the unification of the support and the confidence measures is so wide and there is a lack of consensus about the choice of the appropriate support and confidence measures. In this work, we aimed at providing a unifying definition of support and confidence measures within evidential database. We proposed a new representation of support and confidence measures using interval arithmetic. This representation is bounded by the lower and the upper values that the support (resp. confidence) could take. From a methodological point of view, this paper includes the following *key contributions*: *(i)* definition of new measures of support and confidence expressed with intervals within evidential databases; *(ii)* development of new mining algorithm, named *EBS-Apriori*, that retrieves frequent patterns and valid association rules; *(iii)* implementation of an associative classifier algorithm based on weighted valid association rules and evidence theory fusion techniques.

This paper is organized as follows: in section 2, the state-of-the-art works of evidential data mining are recalled briefly. In section 3, we introduce new bounded support and confidence measures. The EBS-Apriori mining algorithm is detailed in 4. In addition, several strategies for patterns and association rules pruning are presented. The performance of our proposed approach was studied on several database benchmarks in section 5. Finally, we conclude and sketch potential issues for the future work.

2 Preliminaries: Evidential data mining

In this section, we present briefly the main concepts of data mining over evidential databases.

Definition 1. *Evidential databases [8] aim at handling imprecise and uncertain data. Formally, an evidential database is a triplet* $\mathcal{EDB} = (\mathcal{A}_{\mathcal{EDB}}, \mathcal{O}, R_{\mathcal{EDB}})$. $\mathcal{A}_{\mathcal{EDB}}$ *and* $\mathcal{O}$ *are respectively the set attributes and d transactions (i.e., rows). Each column* A_i $(1 \leq i \leq n)$ *has a domain* Θ_i *of discrete values.* $R_{\mathcal{EDB}}$ *expresses the relationship between the* j^{th} *transaction (i.e., row* T_j*) and the* i^{th} *column (i.e., attribute* A_i*) by a normalized BBA* $m_{ij} : 2^{\Theta_i} \rightarrow [0, 1]$ *as follows:*

$$\begin{cases} m_{ij}(\emptyset) = 0 \\ \sum\limits_{\omega \subseteq \Theta_i} m_{ij}(\omega) = 1. \end{cases} \tag{1}$$

Table 1 illustrates an example of an evidential database. An item corresponds to a focal element[1]. An *evidential association rule* R is a causal relationship between two itemsets that

[1]Each subset A of 2^{Θ}, fulfilling $m(A) > 0$, is called a focal element.

can be written in the following form $R : X \rightarrow Y$ such that $X \cap Y = \emptyset$. Two different itemsets can be related via either the inclusion or the intersection operator. Indeed, the inclusion operator for evidential itemsets [6] is defined as follows, where X and Y are two evidential itemsets: $X \subseteq Y \iff \forall x_i \in X, x_i \subseteq y_j$.

x_i and y_j are respectively the i^{th} and the j^{th} element of X and Y. For the same evidential itemsets X and Y, the intersection operator is defined as follows: $X \cap Y = Z \iff \forall z_k \in Z, z_k \subseteq x_i$ and $z_k \subseteq y_j$.

Example 1. *In Table 1, A_1 is an item and $\Theta_A \times B_1$ is an itemset such that $A_1 \subset \Theta_A \times B_1$ and $A_1 \cap \Theta_A \times B_1 = A_1$. $A_1 \rightarrow B_1$ is an evidential association rule.*

As it is the case for probabilistic data mining [14], the support within the evidential context is based on expectation. Two support family approaches were proposed. The first support measure was proposed by [6] and called the belief-based support measure. It is considered as the lower bound for the support. It is written as follows:

$$Sup^{Bel}_{T_j}(X) = \prod_{i \in [1 \ldots n]} Sup^{Bel}_{T_j}(x_i) = \prod_{i \in [1 \ldots n]} Bel(x_i) \tag{2}$$

such as the belief function $Bel(.)$ is computed as $Bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B)$. Thus, the belief-based support in the entire database is computed as follows:

$$Sup^{Bel}_{\mathcal{EDB}}(X) = \frac{1}{d} \sum_{j=1}^{d} Sup^{Bel}_{T_j}(X) \tag{3}$$

Since the belief-based support is a lower estimation of the support, it is obvious in some cases that an itemset I could have a higher support value. Another measure was introduced by Samet et al.[13] that provides a medium estimation. The evidential support of an itemset $X = \prod_{i \in [1 \ldots n]} x_i$ in the transaction T_j (i.e., Pr_{T_j}) is then computed as follows:

$$Sup^{Pr}_{T_j}(X) = \prod_{x_i \in \Theta_i, i \in [1 \ldots n]} \sum_{x \subseteq \Theta_i} \frac{|x_i \cap x|}{|x|} \times m_{ij}(x) \qquad \forall x_i \in 2^{\Theta_i}. \tag{4}$$

Thus, the evidential support $Sup^{Pr}_{\mathcal{EDB}}$ of the itemset X becomes:

$$Sup^{Pr}_{\mathcal{EDB}}(X) = \frac{1}{d} \sum_{j=1}^{d} Sup^{Pr}_{T_j}(X). \tag{5}$$

A new metric for confidence computing based on the precise-based support measure is introduced in [13]. For an association rule $R : R_a \rightarrow R_c$, the confidence is computed as follows:

$$Conf(R) = \frac{\sum_{j=1}^{d} Sup_{T_j}(R_a) \times Sup_{T_j}(R_c)}{\sum_{j=1}^{d} Sup_{T_j}(R_a)}. \tag{6}$$

The precise-based support provides several limits. In fact, computing the support by integrating the disjunction of hypotheses (e.g Θ_i) could lead to incoherent behaviour. Example 2 details the limits of both support measures.

Example 2. *Let us assume the evidential database depicted in Table 1. We aim at computing the support of A_1. The belief-based support gives a support equal to $\frac{0.7}{2}$. This support value is*

Figure 1: The support boundary of $A_1 \times B_1$ in $\mathcal{EDB}$ transaction T_1

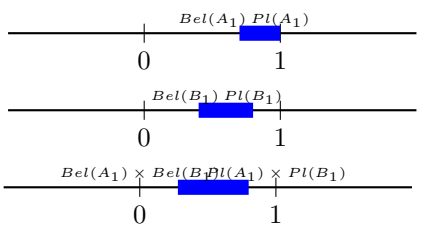
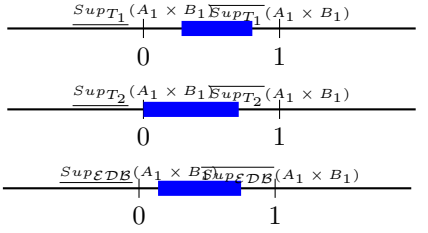


Figure 2: The support boundary of $A_1 \times B_1$ in $\mathcal{EDB}$



the lowest value that A_1 could have. On the other hand, the precise support gives $\frac{0.85+0.35}{2}$ as A_1 support since it belongs by half to Θ_A. Even so, no one can be sure if the support would be equal to those values.

In the following section, we aim at avoiding the limits of state-of-the-art support measure shown in Example 2. We intend to model the support within an interval rather with a single value. Interval support modelling ensures an accurate support. Indeed, with such modelling we are sure that the real value of the support belongs to the retained interval.

3 Bounded support and confidence in evidential databases

In this section, we introduce a bounded value of the support within evidential databases. Thus, we are sure that the real value of the support belongs to the interval. Such methodology can be further extended to probabilistic and fuzzy cases.

The aim is to compute an upper and a lower bound for the $Sup(x)$ that we denote respectively as $\overline{Sup}(x)$ and $\underline{Sup}(x)$. In evidence theory, the belief function $Bel(.)$ in a subset $A \subseteq \Theta$ is interpreted as the belief one actually commits to A. On the other hand, another measure called plausibility $Pl(.)$ is interpreted as the maximum possible belief one may commit to A and is written as follows:

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - Bel(\neg A). \tag{7}$$

We can easily verify that $Pl(A) \geq Bel(A)$. Then, the interval [Bel(A),Pl(A)] represents the quantitative judgements on a proposition A based on a given evidence. Both functions are sometimes referred to, respectively, as the lower and upper probability measures [5]. In addition, the precise support which is an extension of the pignistic probability to the powerset (a.k.a 2^{Θ}) belong to the aforementioned interval.

Definition 2. *Let us assume an itemset $X = x_1 \times \ldots \times x_k$ in the evidential database $\mathcal{EDB}$, the transactional bounded support of X within a transaction j is computed as follows:*

$$\tilde{Sup}_{T_j}(X) = [\underline{Sup_{T_j}}(X), \overline{Sup_{T_j}}(X)] = [\prod_{1 \leq i \leq k} Bel(x_i), \prod_{1 \leq i \leq k} Pl(x_i)]. \tag{8}$$

Definition 3. *Assuming an itemset X within the evidential database $\mathcal{EDB}$, The bounded support of X within the database is computed as follows:*

$$\tilde{Sup}_{\mathcal{EDB}}(X) = [\frac{\sum_{1 \leq j \leq |D|} \underline{Sup_{T_j}}(X)}{|D|}, \frac{\sum_{1 \leq j \leq |D|} \overline{Sup_{T_j}}(X)}{|D|}]. \tag{9}$$

Example 3. *Let us assume the evidential database depicted in Table 1. Figure 1 and 2 illustrate the transactional bounded and the bounded support of the itemset* $A_1 \times B_1$.

Property 1. *The bounded support is interval-wise anti-monotonic. For two itemset* A *and* $A \times X$ *in* $\mathcal{EDB}$*, we have:*

$$\underline{Sup_{\mathcal{EDB}}}(A) \leq \underline{Sup_{\mathcal{EDB}}}(A \times X) \tag{10}$$

$$\overline{Sup_{\mathcal{EDB}}}(A) \leq \overline{Sup_{\mathcal{EDB}}}(A \times X). \tag{11}$$

Proof. Property 1 is proved by reasoning under the constraint $Pl(X) \in [0,1]$ for an itemset $A \times X$. The same goes for the lower bound of the support with the belief function such as $\underline{Sup_{\mathcal{EDB}}}(A \times X) \leq \underline{Sup_{\mathcal{EDB}}}(A)$ [6].

□

Let us assume an association rule $R : R_a \rightarrow R_c$, such that R_c and R_a are respectively the conclusion and the premise part of the rule R. As originally introduced in binary databases, the confidence measure was relying on conditional probability [2]. The confidence could be relying on a probability, fuzzy or an evidential conditional measure depending on the used uncertain framework. The same issues assigned to the state-of-the-art support measures are still valid for confidence measures. Therefore, in the following, we introduce a bounded computation for the confidence. Then, the bounded confidence can be written as follows:

$$\tilde{Conf}(R) = [\underline{Conf}(R), \overline{Conf}(R)] = [\frac{\sum_{j=1}^{d} \underline{Sup_{T_j}}(R_a) \times \underline{Sup_{T_j}}(R_c)}{\sum_{j=1}^{d} \underline{Sup_{T_j}}(R_a)}, \frac{\sum_{j=1}^{d} \overline{Sup_{T_j}}(R_a) \times \overline{Sup_{T_j}}(R_c)}{\sum_{j=1}^{d} \overline{Sup_{T_j}}(R_a)}] \tag{12}$$

Example 4. *Let us assume the evidential database of Table 1. The bounded support of the itemset* $A_1 \times B_1$ *is computed as follows:* $\tilde{Sup}_{\mathcal{EDB}}(A_1 \times B_1) = [\frac{0.7 \times 0.4 + 0}{2}, \frac{1 \times 0.8 + 0.7 \times 1}{2}] = [0.14, 0.75]$. *Then, the confidence of the association rule* $R : A_1 \rightarrow B_1$ *becomes:*

$$\tilde{Conf}(R) = [\frac{0.7 \times 0.4 + 0}{0.7 + 0}, \frac{1 \times 0.8 + 0.7 \times 1}{1 + 0.7}] = [0.4, 0.88].$$

4 Data mining with bounded support and confidence

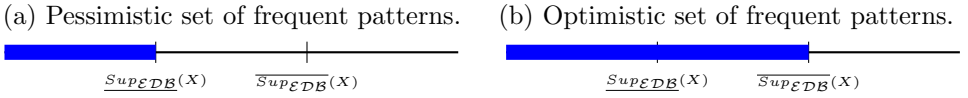
In this section, we investigate the mining process of frequent patterns and valid association rules under the new bounded support and confidence.

4.1 Frequent patterns and valid association rules extraction

We assume that *minsup* and *minconf* (denoted respectively α and β) are two thresholds fixed by the user. Patterns are called frequent if their support are greater than or equal to α. The same goes for association rules which must have a confidence greater than or equal to β to be considered as valid. Now, assuming an itemset I with a support $[a, b]$ and α, under which conditions this pattern is considered as frequent?

Definition 4. *Let* I *be a pattern extracted from the evidential database* $\mathcal{EDB}$ *with a support* $[\underline{Sup_{\mathcal{EDB}}}(I), \overline{Sup_{\mathcal{EDB}}}(I)]$ *and* α *is the support threshold.* I *is a frequent pattern if* $\underline{Sup_{\mathcal{EDB}}}(I) \geq \alpha$. I *is a non frequent pattern if* $\overline{Sup_{\mathcal{EDB}}}(I) < \alpha$.

Figure 3: Range of α for both pessimistic and optimistic pruning strategies



The decision becomes more complex when I has an $\alpha \in [\underline{Sup}_{\mathcal{EDB}}(I), \overline{Sup}_{\mathcal{EDB}}(I)]$. Therefore, in this work, we distinguish three strategies to consider those specific itemsets. The *optimistic approach* consists in considering an itemset frequent as long as α is lower than the upper bound of the support interval (i.e., $\overline{Sup}_{\mathcal{EDB}}(I) \geq \alpha$). The *pessimistic approach* prunes every itemset having an α strictly lower than the lower bound of the support (i.e., $\underline{Sup}_{\mathcal{EDB}}(I) \geq \alpha$). The same methodology could be applied for association rules with β as threshold. In fact, an association rule R is valid as long as β is lower than or equal to the lower bound of R's confidence interval. On the other hand, R is said a non valid rule if β is higher than the upper bound of its confidence interval. Two main strategies for pruning association rules can be distinguished: the *optimistic* and the *pessimistic* strategies. The optimistic strategy retains rules having β lower than or equal to the upper bound of the confidence. On the counter part, the pessimistic strategy consists in retaining any association rule that has a lower bound of the confidence greater than or equal to β. Figure 3 shows the range of α (resp. β for the confidence) for both optimistic and pessimistic strategies.

To mine frequent patterns and valid association rules from evidential databases with the bounded support, a specific EBS-Apriori algorithm is developed. The proposed algorithm is an Apriori-based one [2]. The development of an Apriori-based algorithm is justified by its perfomance over tree-based ones in dense databases [14]. The evidential databases are naturally dense. The algorithm 1 details EBS-Apriori. It is a level-wise algorithm similar to UApriori [4] and the original binary Apriori [2]. The generated candidates are pruned with respect to their computed support. The main difference is that the support is computed as an interval. As the UApriori, EBS-Apriori includes a trimming part [4]. The basic idea behind it is to trim away items with low existential presence from the evidential database and then to mine the trimmed structure. As a result, a structure called $Trim\_Table$ is constructed that stores the belief values (i.e., Bel(.)) of interesting items. The plausibility value is not needed since it can be computed from the belief values (see Equation 7). Depending on the computed support, the itemset is either affected to the set of optimistic frequent patterns (i.e., $\mathcal{OEIFF}$) or pessimistic one (i.e., $\mathcal{PEIFF}$). The bounded support is computed with the use of $Bounded\_Sup(.)$ function that provides two outputs: the lower and the upper bound of the support of a candidate itemset. The support is computed from the $Trim\_Table$ structure. The function $Rule\_generation(.)$ takes as input the set of optimistic frequent itemset (i.e., $\mathcal{OEIFF}$ since $\mathcal{PEIFF} \subseteq \mathcal{OEIFF}$) and generates the set of optimistic and pessimistic valid association rules (i.e., $\mathcal{R}_{opt}$ and $\mathcal{R}_{pes}$).

4.2 Weighted Evidential Associative Classifier

Let us suppose the existence of an instance X to classify represented a set of BBAs belonging to the evidential database $\mathcal{EDB}$ as $X = \{m_i | i \in [1, n]\}$. Each retained association rule, from the rule set $\mathcal{R}$, is considered as a potential piece of knowledge that could be helpful for the class retrieval of X. $\mathcal{R}$ could be either the set of optimistic or pessimistic association rules (i.e., $\mathcal{R}_{opt}$ and $\mathcal{R}_{pes}$). In order to select rules that may lead to the correct classification, we look for association rules having a non null intersection with X and contain a class in the conclusion part, i.e., $\mathcal{RI} = \{R \in \mathcal{R} | \exists x \in \Theta_i, m_i(x) > 0,\ x \in R_a \wedge \exists y \in \Theta_C,\ y \in R_C\}$, where Θ_C is

Algorithm 1 Evidential Bounded Support Apriori (EBS-Apriori)

Require: $\mathcal{EDB}, \alpha, Size\_\mathcal{EDB}, \beta$
Ensure: $\mathcal{OEIFF}, \mathcal{PEIFF}, \mathcal{R}_{opt}, \mathcal{R}_{pes}$
$Trim\_Table \leftarrow construct\_trim(\mathcal{EDB}, \alpha, Size\_\mathcal{EDB})$
$\mathcal{EIFF} \leftarrow \emptyset$, $size \leftarrow 1$
$cand \leftarrow candidate\_apriori\_gen(\mathcal{EDB}, size)$
While$(cand \neq \emptyset)$

for all $X \in cand$ **do**
    $\{u, l\} \leftarrow Bounded\_Sup(cand, Trim\_Table, Size\_\mathcal{EDB})$
    if $l \geq \alpha$ **then**
        $\mathcal{OEIFF} \leftarrow \mathcal{OEIFF} \cup X$
        $\mathcal{PEIFF} \leftarrow \mathcal{PEIFF} \cup X$
    else
        if $u \geq \alpha$ **then**
            $\mathcal{OEIFF} \leftarrow \mathcal{OEIFF} \cup X$
End While
$\{\mathcal{R}_{opt}, \mathcal{R}_{pes}\} \leftarrow Rule\_generation(\mathcal{OEIFF}, Trim\_Table, \beta)$
function RULE_GENERATION$(X, Trim_Table, \beta)$
    for all $x \in X$ **do**
        $R \leftarrow Construct\_Rule(x, \Theta_C)$
        if $R \neq \emptyset$ **then**
            $\{u, l\} \leftarrow Bounded\_Confidence(R, Trim\_Table)$
            if $l \geq \alpha$ **then**
                $\mathcal{R}_{opt} \leftarrow \mathcal{R}_{opt} \cup R$
                $\mathcal{R}_{pes} \leftarrow \mathcal{R}_{pes} \cup R$
            else
                if $u \geq \beta$ **then**
                    $\mathcal{R}_{opt} \leftarrow \mathcal{R}_{opt} \cup R$
    return $\{\mathcal{R}_{opt}, \mathcal{R}_{pes}\}$

the frame of discernment of the class. Each rule found in the set $\mathcal{RI}$ constitutes a piece of information concerning the membership of the instance X. Since several rules can be found and fulfilling the intersection condition, it is important to benefit from them all. In our work, we assume that all information are valuable and should be handled as an information fusion problem. From the set of association rules obtained through the use of optimistic or pessimistic strategy, each rule $R_l \subset \mathcal{RI}$, $l \in [1 \dots L]$ and $L < |\mathcal{RI}|$, is transformed into a BBA with respect to the frame of discernment Θ_C (i.e., frame of discernment of the class in R_c) as follows:

$$\begin{cases} {}^{w}m_{R_l}^{\Theta_C}(\{R_c\}) = W(R_l) \times \gamma \times \underline{Conf}(R_l) \\ {}^{w}m_{R_l}^{\Theta_C}(\Theta_C) = 1 - W(R_l) \times \gamma \times \underline{Conf}(R_l) \end{cases} \tag{13}$$

where $\gamma \in [0, 1[$ prevents from having a certain BBA[2] and R_c is the conclusion part of the rule R_l. $W(.)$ is a weight function taking values in [0,1]. It expresses how much a rule R_l will be considered before rules combination. It is fixed to 1 for either optimistic and pessimistic derived set of rules. In this work, we used the lower bound of the confidence to respect the minimum information principle. The L constructed BBA are then fused following Dempster's

[2]A BBA is called a certain BBA when it has one focal element, which is a singleton. It is representative of perfect knowledge and the absolute certainty.

rule of combination [5] as follows:

$$m_{\oplus} = \oplus_{l=1}^{L} {}^{w}m_{R_l}^{\Theta_C}. \tag{14}$$

When the weight function is fixed to 1, Equation 14 combines all association rules with the same consideration. One problem may arise when applying one of those two pruning strategies. The optimistic strategy could be too optimistic by retaining rules with an upper bound confidence close to β. The same goes for the pessimistic strategy that prunes even association rules with β close to the lower bound of the confidence. Those limits could be problematic for an associative classifier that uses rules for prediction and classification. Therefore, in the following, we introduce a distance-based method to weight association rules having β in the confidence interval. A method would be to compute the distance between β and the upper bound of the confidence. The weight function $W(.)$ can be computed as follows:

$$W(R) = 1 - \frac{\overline{Conf}(R) - \beta}{\overline{Conf}(R) - \underline{Conf}(R)} \quad if\ \beta \in [\overline{Conf}(R), \underline{Conf}(R)]. \tag{15}$$

The weight function $W(.)$ would be of help for developing an alternative to the optimistic and pessimistic strategies in considering rules with β in the confidence interval.

Algorithm 2 details the classification process based on the largest premise rules. The function $FILTRATE\_LARGE\_PREMISE(.)$ (line 1) allows to select valid association rules and to retain only those with the largest premise, having an intersection with the instance to classify or to predict X. In fact, the set of the largest premise rules are more precise than those with a shorter one [13]. Function $Weight(.)$ computes the weight of an association rule as detailed in Equation (15). Once found, they are considered as independent sources and are combined (line 5). The function $argmax$ in line 6 allows the retention of the hypothesis that maximizes the pignistic probability [13].

5 Experiments

The algorithms were applied on several real benchmarks transformed into evidential databases [13]. We used two types of benchmarks. The largest databases such as Skin Segmentation (245057 rows, 4 columns and 32 focal elements), KEGG Metabolic Relation Network (53414 rows, 23 columns and 96 focal elements) and MAGIC Gamma Telescope (19020 rows, 11 columns and 44 focal elements) were used to assess the scalability of the mining algorithm. The smallest databases such as Wine (178 rows, 13 columns and 416 focal elements), Vertebral column (310 rows, 6 columns and 192 focal elements), Diabetes (768 rows, 8 columns and 256 focal elements) and Iris (150 rows, 5 columns and 40 focal elements) were tested to assess the accuracy of the classifier. Even if the number of records and columns seems limited, they expand exponentially in imperfect databases. For example, a database of n columns contains $n \times 2^{clus}$ focal elements (i.e., items) after evidential transforming process using Evidential C-Means (ECM) [13]. $clus$ is the number of clusters given as a parameter to ECM.

Figure 4 (a), (b) and (c) show the runtime performances of several algorithms on the largest datasets. In fact, we compared EBS-Apriori in its optimistic and pessimistic versions (EBS-Apriori-Opt and EBS-Apriori-Pes) to BIT [6] which is the tree-based algorithm that uses the belief-based support. We also made a comparison to EDMA [13] which is also an Apriori-based algorithm that compute the support with the precise measure. It is important to notice that EBS-Apriori-Pes produces the same output as BIT but more time consuming since it computes the support as an interval. EBS-Apriori-Opt is the most time consuming since it generates

Algorithm 2 Weighted Evidential Associative Classification (WEvAC) algorithm

Require: $\mathcal{R}, X, \Theta_C$
Ensure: $Class$
$\mathcal{R}_{large} \leftarrow FILTRATE\_LARGE\_PREMISE(\mathcal{R}, X, \Theta_C)$
for all $r \in \mathcal{R}_{large}$ **do**
  $W \leftarrow Weight(r)$
  $m \leftarrow \begin{cases} m(\{r.conclusion\}) = W \times \gamma \times conf(r) \\ m(\Theta_C) = 1 - W \times \gamma \times conf(r) \end{cases}$
  $m_{\oplus} \leftarrow m_{\oplus} \oplus m$
$Class \leftarrow argmax_{H_k \in \Theta_C} BetP(H_k)$
function FILTRATE_LARGE_PREMISE($\mathcal{R}, X, \Theta_C$)
  $max \leftarrow 0$
  for all $r \in \mathcal{R}$ **do**
    if $r.conclusion \in \Theta_C$ & $X \cap r.premise \neq \emptyset$ **then**
      if $size(r.premise) > max$ **then**
        $\mathcal{R}_{large} \leftarrow \{r\}$
        $max \leftarrow size(r.premise)$
      else
        if $size(r.premise) = max$ **then**
          $\mathcal{R}_{large} \leftarrow \mathcal{R}_{large} \cup \{r\}$
  return $\mathcal{R}_{large}$

much more frequent patterns than the other algorithms. For example, in Magic_EDB database EDMA retrieves a peak of 13749 frequent patterns in contrast with EBS-Apriori-Opt that generates 27891 ones. The results also consolidates that EDMA is more expensive runtime-wise than BIT [13]. Figure 4 (d), (f) and (g) compare the number of extracted frequent patterns with optimistic and pessimistic versions of the support to the precise-based support. The belief-based support was not considered in this comparative study since its results match those of the pessimistic one. The results show a high number of frequent patterns for EBS-Apriori-Opt for all considered databases. The frequent patterns retrieved by the EDMA outnumber those of the EBS-Apriori-Pes (aka belief-based support). As it is the case for binary data mining, a frequent itemset I_k of size k generates $2^k - 2$ association rules. Thus, the number of valid association rules for EBS-Apriori-Opt outnumber all of other approaches. The accuracies of the evidential associative classifiers are depicted in Table 2. We compared WEvAC in its optimistic, pessimistic (i.e., optimistic and pessimistic strategies of association rule's pruning with a fixed weight parameter $W = 1$) versions (WEvAC-Opt, WEvAC-Pes) to the weighted versions (WEvAC). Moreover, we confronted the performance of our presented approaches to the state-of-the-art ones: EDMA classifier [13] and CMAR associative classifier [10]. The results showed that a great number of rules could hamper the accuracy of the classifier as it is the case for WEvAC-Opt. This behaviour can be explained by the outcome of the combination operator within the algorithm. In fact, the more you combine rules, the low accuracy you get. The best results are those of WEvAC. This can be explained by the weight function that adjust the number of used rules and reduces the impact of some rules in the combination stage. EDMA in its associative classifier version provides interesting results similar to those of WEvAC. In addition, our results showed also that WEvAC in its pessimistic and weighted versions outperform the associative classifier CMAR. In fact, the imprecision handling with evidence theory allows us to handle new type of rules with a composed items in the premise part. For example, in wine database, we may have rule: *if you have an Ash between 1.75 and 2.10 then the wine belongs to class 1.* Such kind of rules could improve the accuracy results. Finally, we compared WEvAC to the Support Vector Machine and Neural Networks implemented in

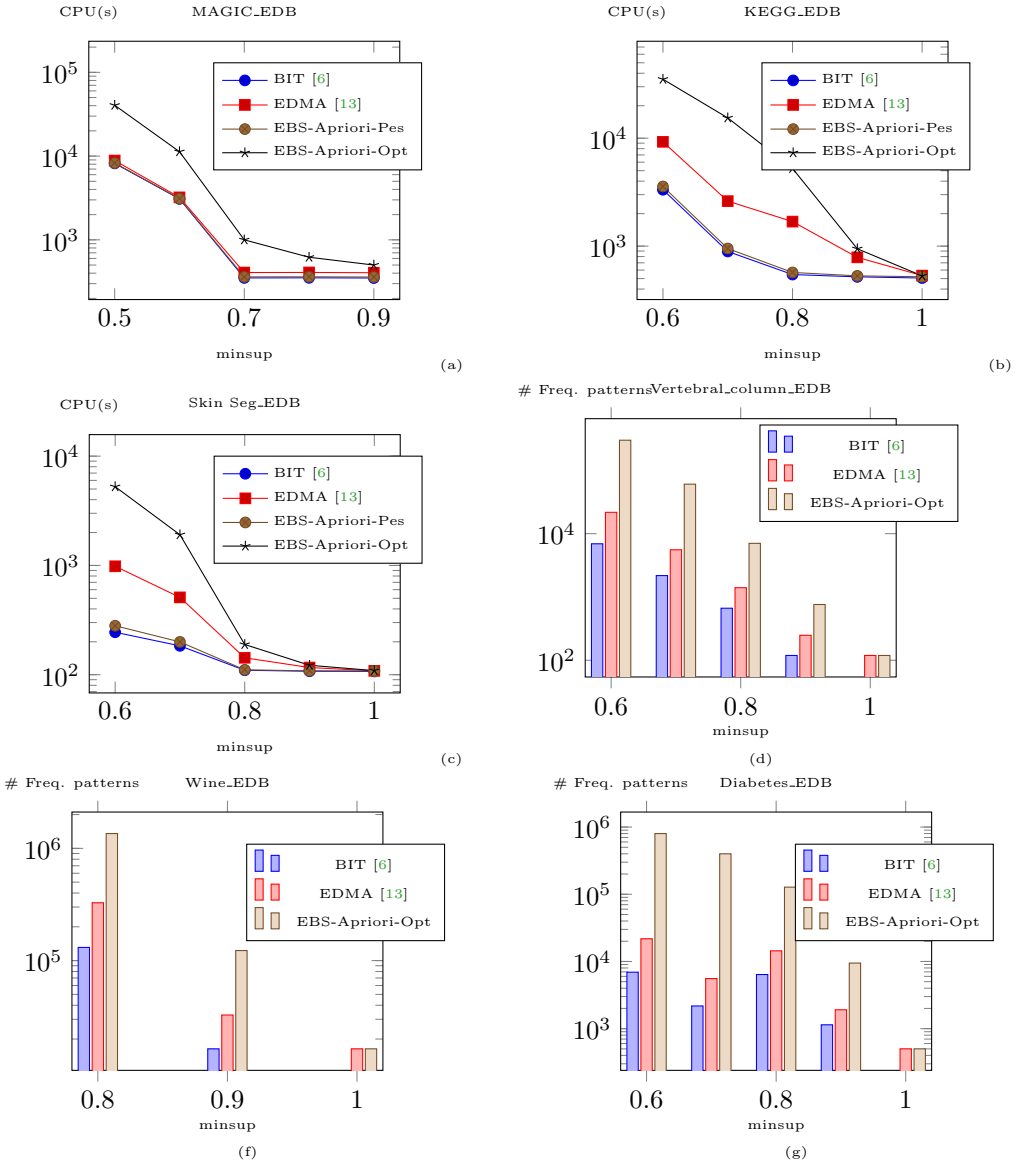


Figure 4: Evaluation results

WEKA software. The results showed the superiority of WEvAC for all databases in terms of accuracy. Those results confirm the impact of imprecision handling within databases.

6 Conclusion

In this paper, we introduced new measures of support and confidence computed as intervals within the evidential database framework. A new mining algorithm and an associative classifier were also developed and analyzed. As illustrated in the experiment section, the proposed

Dataset	WEvAC-Opt	WEvAC-Pes	WEvAC	EDMA [13]	CMAR [10]	SVM	N. Net
Diabete_EDB	77.37%	79.17%	**85.15%**	83.20%	75.10%	77.47%	80.60%
Wine_EDB	76.40%	91.57%	**100%**	100%	95.00%	99.43%	**100%**
Vertebral column_EDB	67.74%	83.87%	**88.39%**	85.16%	81.61%	80%	87.74%
Iris_EDB	73.33%	79.33%	**82.00%**	80.67%	81.61%	96.00%	97.33%

Table 2: Classification accuracies for several transformed datasets

approach provided an interesting performance on several database benchmarks. In future work, we will be interested in generalizing the proposed approach in probabilistic and fuzzy databases. Furthermore, the performance of EBS-Apriori algorithm could be improved by adding specific heuristics such as the decremental pruning [1].

Acknowledgement

This work was performed, in partnership with the SAS PIVERT, within the frame of the French Institute for the Energy Transition (Institut pour la Transition Energétique (ITE) P.I.V.E.R.T. (www.institut-pivert.com) selected as an Investment for the Future ("Investissements d'Avenir"). This work was supported, as part of the Investments for the Future, by the French Government under the reference ANR-001-01.

References

[1] C-C Aggarwal. *Managing and Mining Uncertain Data*, volume 3. Springer, 2010.

[2] R. Agrawal and R. Srikant. Fast algorithm for mining association rules. *In Proceedings of international conference on Very Large DataBases, VLDB, Santiago de Chile, Chile*, pages 487–499, 1994.

[3] Y-L. Chen and C-H. Weng. Mining association rules from imprecise ordinal data. *Fuzzy Set Syst*, 159(4):460–474, 2008.

[4] C-K Chui, B. Kao, and E. Hung. Mining frequent itemsets from uncertain data. *in Proceedings of the 11th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Nanjing, China*, pages 47–58, 2007.

[5] A.P. Dempster. *Upper and lower probabilities induced by multivalued mapping.* AMS-38, 1967.

[6] K.K. Rohitha Hewawasam, K. Premaratne, and M-L Shyu. Rule mining and classification in a situation assessment application: A belief-theoretic approach for handling data imperfections. *Trans. Sys. Man Cyber. Part B*, 37(6):1446–1459, 2007.

[7] T-P Hong, M-J Chiang, S-L Wang, et al. Fuzzy weighted data mining from quantitative transactions with linguistic minimum supports and confidences. *International Journal of Fuzzy Systems*, 8(4):173–182, 2006.

[8] S.K. Lee. Imprecise and uncertain information in databases: an evidential approach. *In Proceedings of Eighth International Conference on Data Engineering, Tempe, AZ*, pages 614–621, 1992.

[9] C.K.-S. Leung, R.K. MacKinnon, and S.K. Tanbeer. Fast algorithms for frequent itemset mining from uncertain data. *In Proceeding of IEEE International Conference on Data Mining (ICDM), Shenzhen, China*, pages 893–898, Dec 2014.

[10] W. Li, J. Han, and J. Pei. CMAR: Accurate and efficient classification based on multiple class-association rules. *in Proceedings of IEEE International Conference on Data Mining (ICDM01), San Jose, CA, IEEE Computer Society*, pages 369–376, 2001.

[11] A. Samet, E. Lefevre, and S. Ben Yahia. Classification with evidential associative rules. *In Proceedings of 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Montpellier, France*, pages 25–35, 2014.

[12] A. Samet, E. Lefevre, and S. Ben Yahia. Evidential database: a new generalization of databases? *In Proceedings of 3rd International Conference on Belief Functions, Belief 2014, Oxford, UK*, pages 105–114, 2014.

[13] A. Samet, E. Lefèvre, and S. Ben Yahia. Evidential data mining: precise support and confidence. *Journal of Intelligent Information Systems*, pages 1–29, 2016.

[14] Y. Tong, L. Chen, Y. Cheng, and P-S Yu. Mining frequent itemsets over uncertain databases. *In Proceedings of the VLDB Endowment*, 5(11):1650–1661, 2012.